1. I chose ml.t2.medium instance type for notebook instance because it has 2 vCPUs and 4 GiB of memory, it's more than enough for my needs.
2. I chose t2.micro instance type for EC2 instance because it's comparably cheap and it's enough for my needs.
3. Steps for writing a lambda function:
    a. At first we define the endpoint name we deployed.
    b. Get boto3 client using boto3 session.
    c. Invoke the endpoint using boto3 client by the endpoint name with the image url received from the request.
    d. Process the received result from the endpoint.
    e. Return the processed result.
4. I attached the AmazonGageMakerFullAccess policy to the lambda function's role. I think the attached policy is too permissive but I couldn't find a special policy for invoking an endpoint.
5.
    a. I use 'Reserved Concurrency' and set up 3 maximum concurrent instances for the lambda function. I use 'Reserved Concurrency' because there's no charge for configuring reserved concurrency for a function.
    b. I use these parameters for auto-scaling:
        i. Maximum instance count: 5
        ii. Scale in cool down: 60
        iii. Scale out cool down: 90
        iv. Target value: 7

        I use 60 as a 'Scale in cool down', 90 as a 'Scale out cool down' and 7 as a 'Target value' to make it responsive if simultaneous lambda requests are increased.